Title

Identifying and Predicting Novelty in Microbiome Studies.

Journal

mBio, 9(6)

ISSN

2150-7511

Authors

Su, Xiaoquan
Jing, Gongchao
McDonald, Daniel
et al.

Publication Date

2018-11-01

DOI

10.1128/mbio.02099-18

Peer reviewed

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Applied and Environmental Science

Identifying and Predicting Novelty in Microbiome Studies

Xiaoquan Su,[a,f,g] Gongchao Jing,[a,f,g] Daniel McDonald,[b] Honglei Wang,[a,f,g] Zengbin Wang,[a,f,g] Antonio Gonzalez,[b] Zheng Sun,[a,f,g] Shi Huang,[a,f,g] Jose Navas,[c] Rob Knight,[b,c,d,e] Jian Xu[a,f,g]

[a]Single-Cell Center, CAS Key Laboratory of Biofuels and Shandong Key Laboratory of Energy Genetics, Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, Shandong, China

[b]Department of Pediatrics, University of California San Diego, La Jolla, California, USA

[c]Department of Computer Science & Engineering, University of California San Diego, La Jolla, California, USA

[d]Department of Bioengineering, University of California San Diego, La Jolla, California, USA

[e]Center for Microbiome Innovation, University of California San Diego, La Jolla, California, USA

[f]Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, Shandong, China

[g]University of Chinese Academy of Sciences, Beijing, China

ABSTRACT With the expansion of microbiome sequencing globally, a key challenge is to relate new microbiome samples to the existing space of microbiome samples. Here, we present Microbiome Search Engine (MSE), which enables the rapid search of query microbiome samples against a large, well-curated reference microbiome database organized by taxonomic similarity at the whole-microbiome level. Tracking the microbiome novelty score (MNS) over 8 years of microbiome depositions based on searching in more than 100,000 global 16S rRNA gene amplicon samples, we detected that the structural novelty of human microbiomes is approaching saturation and likely bounded, whereas that in environmental habitats remains 5 times higher. Via the microbiome focus index (MFI), which is derived from the MNS and microbiome attention score (MAS), we objectively track and compare the structural-novelty and attracted-attention scores of individual microbiome samples and projects, and we predict future trends in the field. For example, marine and indoor environments and mother-baby interactions are likely to receive disproportionate additional attention based on recent trends. Therefore, MNS, MAS, and MFI are proposed "alt-metrics" for evaluating a microbiome project or prospective developments in the microbiome field, both of which are done in the context of existing microbiome big data.

IMPORTANCE We introduce two concepts to quantify the novelty of a microbiome. The first, the microbiome novelty score (MNS), allows identification of microbiomes that are especially different from what is already sequenced. The second, the microbiome attention score (MAS), allows identification of microbiomes that have many close neighbors, implying that considerable scientific attention is devoted to their study. By computing a microbiome focus index based on the MNS and MAS, we objectively track and compare the novelty and attention scores of individual microbiome samples and projects over time and predict future trends in the field; i.e., we work toward yielding fundamentally new microbiomes rather than filling in the details. Therefore, MNS, MAS, and MFI can serve as "alt-metrics" for evaluating a microbiome project or prospective developments in the microbiome field, both of which are done in the context of existing microbiome big data.

KEYWORDS microbiome, search, novelty, data mining, bioinformatics, community similarity, database search, microbial ecology, microbiome, microbiome novelty

Received 24 September 2018 **Accepted** 3 October 2018 **Published** 13 November 2018

Citation Su X, Jing G, McDonald D, Wang H, Wang Z, Gonzalez A, Sun Z, Huang S, Navas J, Knight R, Xu J. 2018. Identifying and predicting novelty in microbiome studies. mBio 9:e02099-18. https://doi.org/10.1128/mBio.02099-18.

Editor Margaret J. McFall-Ngai, University of Hawaii at Manoa

Address correspondence to Xiaoquan Su, suxq@qibebt.ac.cn, Rob Knight, robknight@ucsd.edu, or Jian Xu, xujian@qibebt.ac.cn.

This article is a direct contribution from a Fellow of the American Academy of Microbiology. Solicited external reviewers: Edward Ruby, University of Hawaii at Manoa; Emiley Eloe-Fadrosh, DOE Joint Genome Institute.

With the rapid expansion of microbiome sequencing projects around the globe, relating new data to existing data has become one of the most critical bottlenecks for new studies. High-speed comparison and searching for sample similarities in microbiome data sets have been hindered by the lack of appropriate methods. Well-known analytic platforms, such as mothur (1) and QIIME (2), are optimized to support individual projects but not comparisons and searches across all known microbiomes.

Here, we introduce the Microbiome Search Engine (MSE), which, based on taxonomic similarities, rapidly and precisely identifies for each new microbiome sample the best matches from the extremely large number of known microbiomes. MSE consists of two core modules: a well-organized and regularly updated reference database of microbiomes (the entire Qiita public database [https://qiita.ucsd.edu/], which includes 101,983 curated microbiome samples produced by 293 studies between 2005 and 2017) (Fig. S1 and S2; see also Materials and Methods) and a kernel search algorithm (3, 4) (Fig. S3 and S4; see also Materials and Methods). By generating a real-time, landscape-like view of global microbiome compositions from 16S rRNA amplicon data, MSE provides a readily expandable, generally applicable, and widely assessable approach for knowledge-based microbiome analysis.

Tracking the microbiome novelty score (MNS), a metric defined herein based on searching samples against the entire reference database, we detected weak correlation between novelty and alpha-diversity (Spearman $r < 0.4$). Using this metric, we showed that the structural novelty of the human microbiome is approaching saturation and likely bounded, whereas novelty in environmental habitats remains substantially higher. The microbiome focus index (MFI), derived from the MNS and a microbiome attention score (MAS), can objectively track and compare the structural novelty and received attention scores of individual microbiomes or projects and predict trends in the field. For example, marine and indoor environments and mother-baby interactions could be considered "sleeping beauties" soon to be awakened.

RESULTS

Identifying microbiomes with novelty and attention. (i) MNS. By placing each microbiome sample generated so far in the context of the known microbiome space, MSE provides a bird's-eye view of the historical development of global microbiome sequencing efforts. We used all 101,983 curated samples to trace the development of microbiome studies captured in the data set from 2010 to 2017 (because the number of samples began to increase rapidly in 2010). The microbiome novelty score (MNS) was proposed to evaluate the compositional uniqueness of a microbiome sample (at the time of its birth) compared to all microbiomes in the database (see Fig. S5 in the supplemental material). With a given sample, m, and its top n matches, for its match i, whose microbiome similarity is S_i, the MNS(m) was calculated as indicated below (via Meta-Storms [4] similarity of the top 10 matches [see Materials and Methods]).

$$\mathrm{MNS} = 1 - \frac{\sum_{i=1}^{n} \left[S_i \times (n - i) \right]}{\sum_{i=1}^{n} (n - i)} \quad (1)$$

For each microbiome sample, its MNS was derived by searching its sequence against those of all samples produced by past studies (e.g., for a sample published in 2012, its MNS was computed based on its similarity to samples produced prior to 2012). Thus, a higher MNS means lower similarity to those microbiomes that have previously been sampled, suggesting higher novelty. MNS generally followed the normal distribution (Pearson $r = 0.92 \pm 0.07$; two-tailed t test P value $= 0.98 \pm 0.02$, no significant difference [P value ≥ 0.01] compared to a simulated normal distribution) (Fig. 1A), suggesting that the number of samples was adequate. The mean of this distribution in the first year of 2010, 0.15, was chosen as the baseline, and samples with an MNS of $\geq$0.15 were considered novel.

The annual pattern of MNS variation revealed that, although the number of microbiome samples had increased rapidly (there was up to a 36-fold increase from 2010 to

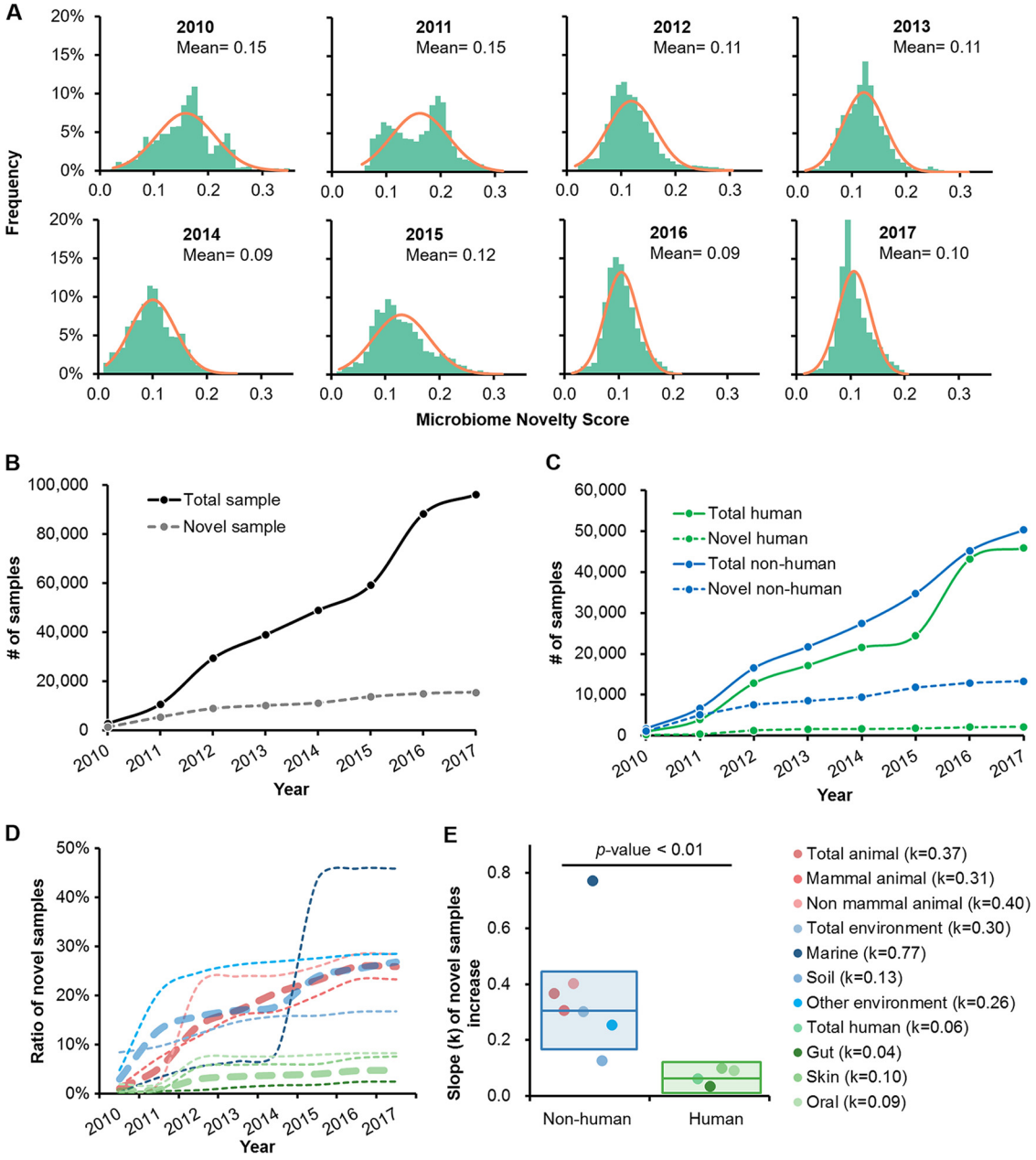

FIG 1 Historical trend of microbiome novelty scores. (A) The MNSs of samples from 2010 to 2017 followed a normal distribution. In each subpanel, the bar chart represents the frequencies of samples and the curve is the simulated standard normal distribution. (B) Yearly accumulative curves of the total numbers of samples and novel samples. From 2010 to 2017, 15,501 samples were identified as novel microbiomes with an MNS of ≥0.15. (C) Yearly accumulative curves of sample numbers for human samples and nonhuman sample. (D) Yearly development of novel sample ratios (defined as the number of novel samples over the number of total samples) in each category. Thick dotted lines represent the ratios of novel samples in high-level categories (human, animal, and natural environments), while thin dotted lines are those in subcategories. (E) Linearly fitting slopes of novel sample ratio increases in each category. The color schemes are the same for panels D and E.

2017), the increase of novelty was much slower (only 10-fold over the same period) (Fig. 1B). In fact, from 2010 to 2017, population-scale studies have continued to resample microbiomes from certain habitats, such as human body sites, causing the unidirectional reduction of mean MNS each year (Fig. 1A). This temporal pattern indicates that current strategies for expanding the boundary of known microbiota is decreasingly efficient, and a new strategy may be required. It is also possible that the natural variation of microbiome compositions is bounded and that the diversity sampled might be approaching saturation.

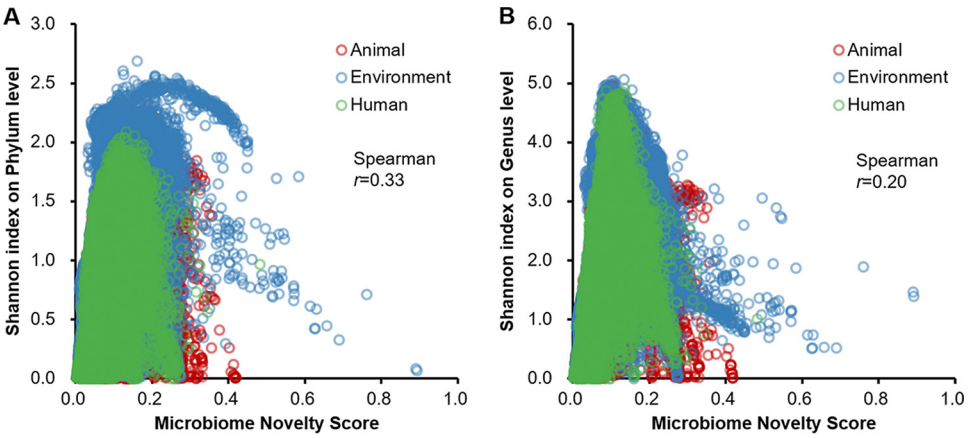

FIG 2 Lack of correlation between the MNSs and Shannon indexes of alpha diversities at both the phylum level (A) and the genus level (B).

However, the relationships between sample volume and number of novel samples can vary widely among ecosystems. For example, although the total number of human microbiomes and nonhuman microbiomes were roughly equivalent ($n = 45{,}813$ versus $n = 50{,}268$), there were 5-fold-more novel samples from nonhuman habitats than human-derived samples (13,329 versus 2,172) (Fig. 1C). Comparison of the trends of novel samples in each subcategory revealed significantly lower linearly fitting slopes of novel sample ratios (defined as the number of novel samples divided by the number of total samples) (Fig. 1D and Fig. S6) for human samples than for nonhuman samples (two-tailed *t* test *P* value < 0.01) (Fig. 1E). Thus, many more previously unknown microbiome compositions are from environmental habitats than from human-associated ones. Among environmental microbiomes, animal (30.17%, of which 16.75% was mammal contributed), lake (17.97%), marine (15.37%), and soil (10.34%) samples had the most-novel microbiomes; in comparison, among human-associated microbiome compositions, those from the gut, skin, and mouths of humans contributed only 4.07%, 4.71%, and 4.44%, respectively, to the novelty.

In addition, the observed novelty was only weakly associated with the community's compositional complexity (Fig. 2), as indicated by the low Spearman correlation between the MNS and the Shannon index at the levels of both the phylum ($r = 0.33$) (Fig. 2A) and the genus ($r = 0.20$) (Fig. 2B). Furthermore, the MNS was also resistant to variation of amplicon regions of microbiome data, which was verified by the same batch gut samples ($n = 150$) that were amplified from the V1-V3 and V3-V5 regions, respectively (two-tailed *t* test *P* value ≥ 0.01) (Fig. S7 [refer to the supplemental results for details]).

On the other hand, within human habitats, at each of the three major human body sites, the gut, skin, and mouth, the trend in accumulation of novel samples slowed in 2012 and then eventually flattened (slope $k = 0.06$) (Fig. 1D and Fig. S6). Among the various body sites in humans, despite its highest sample volume, the gut contributes the fewest novel samples compared to the mouth and skin (gut, 631 of 25,936 samples, with a *k* equal to 0.04; oral, 688 of 8,365 samples, with a *k* equal to 0.09; skin, 730 of 9,657 samples, with a *k* equal to 0.10), resulting in the lowest rate of gain in novel samples over the last 5 years (Fig. S6). Notably, for either gut, oral, or skin samples or all human-associated samples, such rates started to enter a more flattened phase in 2012, which was due to the influx of samples from the Human Microbiome Project (5) published in the same year. This underscores the broad and dramatic impact of such systematic studies in expanding the boundary of microbiome novelty. In this way, few novel microbiotas (those with an MNS of ≥ 0.15) inside or on the human body remain to be discovered, at least in the host populations that are heavily represented at present.

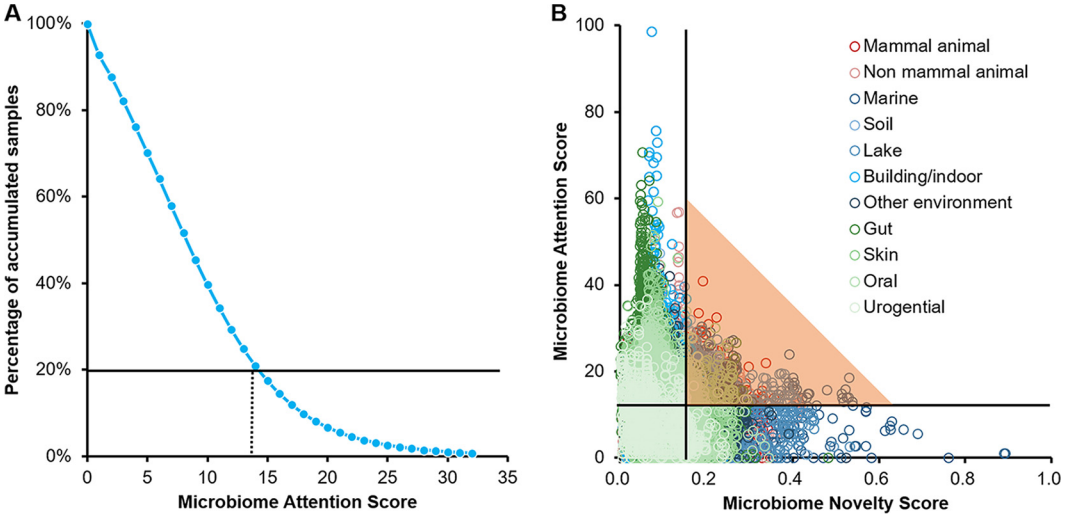

FIG 3 Microbiome attention scores of known microbiome samples. (A) The MAS threshold of 14 is determined based on the top 20% of MAS samples. (B) Distribution of samples by MNS (*x* axis) and MAS (*y* axis). With the cutoff of MNSs was ≥0.15 (novel samples) and that of MASs was ≥14 (high-attention samples), a total number of 2,238 microbiomes were identified as focus samples (dots under the shadows).

(ii) MAS. A parameter based on structural novelty alone is unable to capture the full structural features of a microbiome. A high-MNS sample that spearheads exploration into a new ecosystem can have a high research impact by being subsequently followed by additional sequencing efforts that reveal similar microbiome configurations, or alternatively it can remain "asleep" until such follow-up sequencing ensues. To measure and distinguish such effects, we proposed the microbiome attention score (MAS), which measures the connectivity of a given sample to all subsequent samples in the repository. For a given sample (*m*), its MAS among a total of *n* samples is as follows:

$$\mathrm{MAS} = \sum_{i=0, i\neq m}^{n-1} \mathrm{connectivity}\left(m, i\right) \tag{2}$$

where the connectivity to arbitrary sample *i* [connectivity(*m*, *i*)] is defined as the microbiome similarity between samples *m* and *i* (S_i) if *m* is an element of the top *n* matches of *i* and S_i is ≥0.85, or it is 0 if *m* is an element of the top *n* matches of *i*. In other words, MAS is the similarity sum of samples that match sample *m* with a relative high similarity (Meta-Storms similarity ≥ 0.85) (Fig. S8); hence, a higher MAS indicates that more samples with similarity or samples with higher similarity had been sequenced, suggesting higher attention from the scientific community for this input sample. We also set *n* as 10 for consistency in this work.

To avoid the possible artificial inflation of MAS (and, thus, MFI), such as that caused by redundant sampling from identical microbiotas, we have implemented the following: (i) all reference samples were collected from Qiita, which contains high-quality microbiome studies with extensive metadata; (ii) duplicate samples with a similarity to the existing reference samples of >99.99% were removed from the reference database; and (iii) when calculating MAS, samples from the same study were excluded.

(iii) MFI. We designated the top 20% of the most frequently matched samples by Meta-Storms similarity (corresponding to the threshold MAS of 14) (Fig. 3A) as having high attention among all samples during 2005 to 2017. Hence, samples that have the two attributes of an MNS of ≥0.15 when first sequenced and an MAS of ≥14 were considered "focus" samples (Fig. 3B). A microbiome focus index (MFI), which quantitatively measures the combined novelty and attention of a focus microbiome, is thus calculated as follows:

$$\mathrm{MFI} = \mathrm{MNS} \times \mathrm{MAS} \tag{3}$$

During the 8 years from 2010 to 2017, 2,238 microbiome samples were identified as

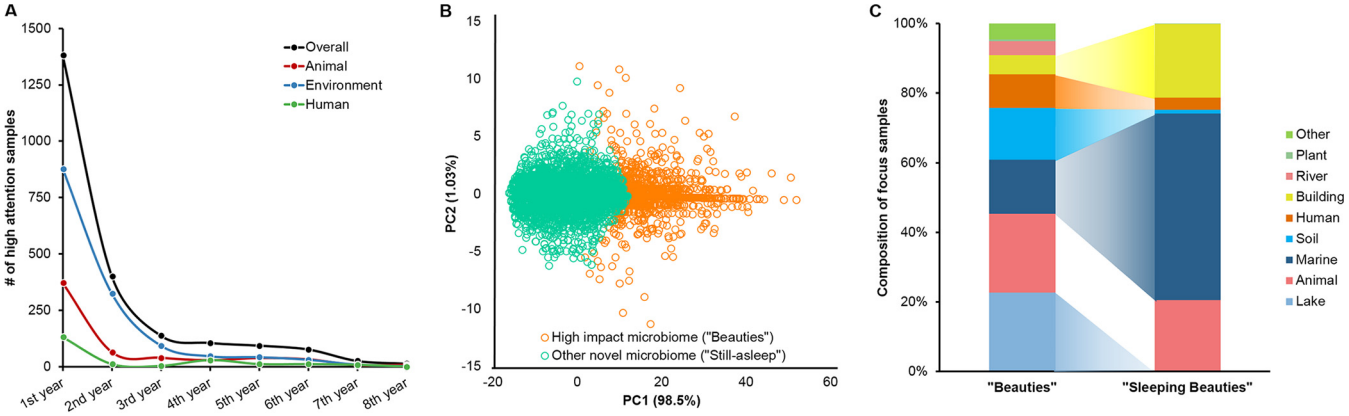

FIG 4 Prediction of sleeping beauty (potential focus) microbiomes. (A) Numbers of focus microbiomes (beauties) that were awaken at the *n*th year after their birth; (B) principal-component analysis of 4-year MASs between beauty samples and still-asleep samples with a random-forest accuracy of 98.78%; (C) habitats of awakened beauties during 2010 to 2017 and of predicted sleeping beauties born since 2015.

having such a focus. The lake (22.29%), animal (22.25%, monkey gut, mouse gut, wild deer gut, dog flea, etc.), marine (15.32%, saline seawater, sponge), soil (14.52%), and human (9.47%, skin, oral, gut, etc.) environments were the environmental types that contributed the most to the highly focused samples. Thus, the MFI derived from the MNS and MAS can serve as a new venue to quantitatively, objectively, and comprehensively evaluate the structural or compositional uniqueness and connectivity of a microbiome among a huge number of samples and studies, which potentially offers advantages to the conventional bibliometric approaches, such as the journal impact factor based on the citation number of the publication, and can be considered an alternative metric (alt-metric) of contribution in the exploration into microbiome space.

Predicting and tracking the focus microbiome. From 2010 to 2017, among the total 15,501 novel microbiomes (with an MNS of ≥0.15), only 2,238 were identified as focus samples by equation 3, while others were nonfocus samples, i.e., present in the repository with few connections (with few structurally similarly microbiomes sequenced and with a low MAS). For the focus microbiomes (with both an MNS of ≥0.15 and an MAS of ≥14), from entry into the database, each of them spends a period awaiting discovery by other researchers (during which they have a low MAS), followed by a process of receiving increasing attention (which increases the MAS). Of particular interest are novel samples (with an MNS of ≥0.15) that still have a low MAS (when there are few structurally similar microbiomes) yet have high potential that will be realized only after a certain amount of time has elapsed, i.e., when a large number of structurally similar microbiomes are deposited and their MASs are therefore increased. On the other hand, most microbiomes may never get high attention. Can we predict these "sleeping beauties" that are currently neglected but will later receive high attention from the ~15,000 microbiomes?

Because the MNS is a constant value set when a sample is first published, the key to predicting potential high-focus samples is to identify whether a novel sample would get high attention (i.e., it has an MAS of ≥14) after its birth year. Thus, using historical data, we asked when a novel sample would receive an MAS above threshold. The yearly development curve of focus samples revealed that most known focus samples (90.6%) garnered attention in their first 4 years (Fig. 4A). This result was confirmed by the first 4 years' MAS pattern of all novel samples produced in 2010 to 2014 (samples after 2015 had only a 3-year MAS); in fact, our random-forest model discriminates focus and nonfocus samples with 98.78% accuracy (Fig. 4B). Based on these results, we built a hybrid model via random-forest regression (see Material and Methods) using the 4-year MASs of novel samples in 2010 to 2014 to predict the sleeping beauties in 2015 to 2017. This model took the <3-year MASs of novel samples as input and estimated their maximum MASs in the future, with a threshold of expected maximum MASs of ≥14 for these potential focus samples.

TABLE 1 Habitats of focus samples, or beauties, during 2010 to 2017 and of predicted potential focus samples, or sleeping beauties, that were born since 2015

Environment	No. of focus samples	No. of predicted focus samples
Lake	499	0
Animal	498	137
Marine	343	358
Soil	325	8
Human	212	23
Building	122	141
River	88	0
Freshwater	39	34
Plant	13	1
Other	99	0

In contrast to the 2,238 focus samples, the 702 potential focus samples predicted to be activated in the next 4 years were found primarily among marine samples (51.00%, saline seawater, sponges, etc.), building and indoor-environment samples (20.09%, glove surface, water heater, etc.), and animal samples (19.52%, monkey gut, horse gut, mouse gut, etc.), while the proportion of focus samples from lakes and soil were significantly reduced (Fig. 4C; Table 1). Thus, based on recent historical trends, we predict that microbiomes with novel community structures from marine and indoor environments are much more likely to be followed up with additional sampling studies and will become hot spots for microbiome research in the near future until they are sampled as completely as human microbiomes.

For human microbiomes, the proportion of focus samples decreased from 9.47% to 3.28%, and only 23 samples were considered sleeping beauties (Fig. 4C), mainly due to the now much rarer high-MNS samples from human microbiomes (as shown in Fig. 1C and D). Sixteen of the 23 potential focus microbiomes were sampled from skin in a mother-baby microbial transfer (6). Thus, on the human front, those focusing on mother-baby interactions are predicted to receive extraordinary attention in the next several years.

Web portal of MSE for computing MNS, MAS, and MFI in real time. To support online microbiome analysis via MSE, a Web portal is provided at http://mse.single-cell.cn/ (registration or login is not required; see Materials and Methods). For the microbiomes in our database, both metadata (e.g., study description, habitat, sequence type, sampling location and date, etc.) and taxonomical structure are provided for online browsing. When users upload a query microbiome in the form of an operational taxonomic unit (OTU) table, the website returns in real time the matched samples from the database, supplemented with their degrees of similarity, their taxonomic compositions, their MNSs, MASs, and MFIs. In addition, the MNS of the query microbiome is also provided. Furthermore, microbiomes that are of similar taxonomic composition to the query can be downloaded for further analysis.

While the MSE reference database is updated regularly, the MNS of a given microbiome in the reference database remained unchanged with time. This is because, per definition, MNS evaluates the compositional uniqueness of a microbiome sample, at the time of its birth, compared to all microbiomes in the database. In contrast, the MAS is dynamic, since the more structurally similar samples emerge, the higher MAS of a microbiome will be, despite its unchanged MNS. As the product of the MNS and MAS, the MFI is also dynamic. Both the MAS and MFI are updated when the reference database is updated.

DISCUSSION

Although an enormous volume of large- and small-scale microbiome data sets from various habitats and produced by different studies have been deposited into public data repositories (e.g., HMP [5], EMP [7], and AGP [8]), there are currently few approaches that scale to process and integrate all the microbiome data so that a global view of microbiomes can be generated in real time (9). This has resulted in the majority of microbiome samples being of single use, that is, that suffer from limited data reuse or citations beyond the original scope of the study. Thus, their value depreciated abruptly.

A search-based strategy, such as MSE, which features a microbiome-composition search accelerated up to 3 orders of magnitude relative to the search capabilities of existing strategies (i.e., pairwise comparisons) in databases of 100,000 to 1,000,000 samples, enables such s bird's-eye view of how each microbiome relates to global microbiomes. For example, by quantitatively defining the novelty of a microbiome, MSE revealed that the novelty of human microbiomes is bounded by their taxonomic dimensions. In fact, efforts expanding the boundaries of the known microbiotas in human (but not nonhuman) habitats has almost reached a plateau, suggesting that a new strategy, such as one focusing on strain- or isolate-level structure or functional variation, might be required.

On the other hand, new metrics, such as the MNS, MAS, and MFI, provide a new way of assessing structural novelty and attention attracted by samples, studies, and areas at a single-microbiome resolution. This quantitative metric, which depends only on the data themselves, may be inherently more accurate and less prone to the influence of unrelated factors than journal or paper citations. Notably, as the definition of the MFI suggests, for focus samples, both the MNS and MAS are important to the MFI. Although no single metric can accurately or thoroughly assess the scientific impact of a microbiome sample, we can argue that focus samples, i.e., those with both high MNSs and high MASs, are likely particularly valuable contributions to our exploration of the microbiome space. For example, by predicting potential focus samples based on the historical evolution of novelty and attention, MSE might potentially help advise policy makers and the scientific community on strategies that efficiently explore the unknown space of microbiome structures. However, appropriate caution should of course be taken against overreliance on any single metric or data source when developing policy.

Finally, MSE is readily expandable, generally applicable and widely assessable. Moreover, because MSE accepts a compositional profile (e.g., OTU, KEGG Orthology, etc.) of a microbiome as search input, the analyses can accommodate both amplicon data sets and metagenomic data sets. We envision that such search against the microbiome database will be an important first step for data analysis at various scales in microbiome studies, just as a BLAST search is essential and universal in sequence analysis studies today.

MATERIALS AND METHODS

MSE reference database. MSE aims to rapidly identify the similar whole-microbiome-level samples of a given query microbiome from a large-scale depository of known microbiomes. The database module, in its present form, consists of the entire Qiita public database (https://qiita.ucsd.edu/), which includes 124,025 public microbiome samples produced by 293 studies (in total) between 2005 to 2017 (see Fig. S1 in the supplemental material), which is regularly updated by adding newly released or published data sets. These studies included the Human Microbiome Project (5), Earth Microbiome Project (7), American Gut Project (8), and other high-impact microbiome studies that cover 18 sampling sources of human-associated habitats, indoor buildings, animal-associated habitats, and various types of natural environments (10–12). Sequences of the 16S rRNA amplicons from the V1-V2, V1-V3, V3-V5, V4, and V6-V9 regions were produced by Illumina HiSeq, MiSeq, or Roche 454 sequencing. After quality control and duplication removal (refer to see "Profiling and normalization" below) (Fig. S2), 101,983 curated microbiome samples were retained for further analysis and interpretation.

Indexing-based search algorithm of MSE. The search module of MSE performs a two-tier indexing process (3, 4), as follows. First is the microbiome feature-based dynamic indexing for fast fetch. The dynamic indexing partitions the OTUs on a specified taxonomy level (also referred to as index keys). Therefore, for each sample, the weight of an index key is the sum of relative abundance values that belong to this index key. When constructing the database, MSE precomputes the index keys and their weights for all database samples (Fig. S3A). Then, for a given query microbiome, MSE calculates its index keys in the same way and dynamically selects candidate matches that have the shortest distances to the query on index keys (Fig. S3B). This reduces the time complexity of searching without the loss of match precision.

The second indexing process is whole-microbiome-level similarity computation with structure reencoding-based optimizations. After indexing, MSE identifies the top matches by a pairwise comparison between the query and each of the candidate matches using the Meta-Storms similarity scoring function (4). This algorithm employs a phylogeny-based metric based on the OTUs' relative abundances to quantitatively assess the similarity between two microbiomes. Typically, the microbiome structures from multiple samples are kept as one centralized file (in BIOM [13], CSV, plaintext, or other equivalent formats) which needs to be entirely loaded into RAM to avoid extra HDD I/O (Hard Disk Drive Input and Output) operations during sample comparison; however, this causes unacceptable memory consumption when 100,000 or more samples are processed (Fig. S4A). To tackle this problem, MSE reencodes the

database microbiome structures by sorting the OTUs by relative memory address offsets in the phylogeny tree, so that the community information is directly placed to the right address. Thus, this MSE approach minimizes the loading cost from HDD (Fig. S4B). Furthermore, MSE separates each reference sample's structure into one individual file that is dispersedly stored in the file system (Fig. S4B). When searching against the whole database, MSE loads only those candidate samples from their particular reencoded files, which maximizes the efficiency of memory usage. Therefore, advanced computing optimization in the two-tier search procedure of MSE greatly and efficiently reduces both the time complexity and the space complexity for large-scale (e.g., over 100,000 samples) microbiome search.

Profiling and normalization. All collected samples from Qiita are profiled and annotated by Parallel-META 3 (version 3.4.2) (14) with Greengenes 13-8 (15) on the OTU similarity level of 97%. Variation in 16S rRNA copy number was normalized based on the IMG/M database (16) to maximally reduce the bias of comparison with samples from different platforms and studies. We set the minimum sequence number to 500 and a minimum 16S rRNA mapping rate of 80% for each sample to ensure high quality of the reference data sets (Fig. S2). We also set a threshold of the Meta-Storms similarity of 99.99% to remove duplicated samples. If the similarity between two samples is equal to or higher than the threshold, the one that has a later production/sampling date in the meta-data is dropped. Finally, 101,983 samples passed the quality control and curation (refer to Data Set S1 for study and sample identification numbers).

Construction of a mixed model for maximum MAS estimation using regression and random-forest modeling. The 4-year MASs of each samples were first normalized by the maximum MAS (MAS_{max}) to compute the maximum MAS ratios (MAS/MAS_{max} is always between 0 and 1), and then a regression model was constructed using the maximum MAS ratios of all novel samples between 2010 and 2014 to describe the 4-year development of attention. In this model, the *x* axis represents the year, while the *y* axis represents the expected maximum MAS ratio calculated by the regression. We also computed the random-forest importance of each year's maximum MAS ratios, and then the maximum MAS of each sample produced after 2015 can be estimated by the following equation:

$$\mathrm{MAS}_{\max} = \frac{\sum_{i=1}^{Y-2014}\left(MAS_i \times \frac{\mathrm{RF}_i}{\mathrm{Reg}_i}\right)}{\sum_{i=1}^{Y-2014}\mathrm{RF}_i} \quad (4)$$

Here, *Y* is the samples' birth year (between 2015 and 2017), MAS_i is the *i*th year's MAS value, Reg_i is the *i*th year's maximum MAS ratio calculated by the regression, and RF*i* is the random-forest importance of the *i*th year's maximum MAS ratio.

Data availability. All samples (including the sequence files and metadata) used in this study are available from Qiita (http://qiita.ucsd.edu). Detailed information about samples that passed the quality control check is provided in Data Set S1 in the supplemental material.

Code availability. MSE is developed and implemented in C/C++. The indexing and searching algorithm is optimized for parallel computing based on multiple CPUs using the OpenMP library. Both source code and executive binary application packages are available at http://mse.single-cell.cn. With this package, users can build their own reference microbiome databases and perform database searches by any given sample. The search results are compatible with Parallel-META 3 software, so the link between the query sample(s) and searching result(s) can be further mined readily (e.g., analyses of taxonomical diversity, the cooccurrence network, and biomarkers). A means to calculate the microbiome novelty score and microbiome attention score is also included in this package.

Website portal and online system. We also provide an online searching engine via a website portal for public use of MSE (http://mse.single-cell.cn). This system accepts input query samples in Parallel-META 3 format and returns search results for both the query sample and matched sample(s) in visualized graphics from multiple perspectives, including a bar chart at the phylum level that shows the comparison of microbiota compositions, the MNS of the query sample derived from all existing database samples, and a result table with similarity values for a matched sample(s) and detailed metadata for in-depth interpretation of the query. A dynamic scheduling strategy is developed and utilized to avoid the task jam for multiple users. In addition, all reference studies and data sets are available for online browsing, and the original sequences and metadata files are also open for download so that user can build their standalone searching environment.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.02099-18.

TEXT S1, DOCX file, 0.01 MB.
FIG S1, JPG file, 0.1 MB.
FIG S2, JPG file, 0.1 MB.
FIG S3, JPG file, 0.2 MB.
FIG S4, JPG file, 0.2 MB.
FIG S5, JPG file, 0.1 MB.
FIG S6, JPG file, 0.5 MB.

FIG S7, JPG file, 0.04 MB.
FIG S8, JPG file, 0.1 MB.
DATA SET S1, XLSX file, 4.8 MB.

ACKNOWLEDGMENTS

J.X. acknowledges support from grants 31327001 and 31425002 from the NSFC and grants KFZD-SW-219-4 and ZDBS-SSW-DQC002-03 from the CAS. X.S. acknowledges support from grants 31771463 from the NSFC, KFZD-SW-219-5 from the CAS, and ZR2017ZB0421 from the Natural Science Foundation of Shandong Province. R.K. acknowledges support from the U.S. National Institutes of Health and the National Science Foundation.

J.X., R.K., and X.S. conceived the idea. X.S. and G.J. developed the software and algorithm. X.S., D.M., H.W., Z.S., and S.H. performed analyses. A.G., J.N., H.W., and Z.W. contributed to data collection and curation. G.J. and Z.W. developed the website portal and online system. X.S., J.X., and R.K. wrote the manuscript.

We declare that we have no competing interests.

REFERENCES

1. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541. https://doi.org/10.1128/AEM.01541-09.
2. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336. https://doi.org/10.1038/nmeth.f.303.
3. Su X, Xu J, Ning K. 2012. Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. Bioinformatics 28:2493–2501. https://doi.org/10.1093/bioinformatics/bts470.
4. Su X, Wang X, Jing G, Ning K. 2014. GPU-Meta-Storms: computing the structure similarities among massive amount of microbial community samples using GPU. Bioinformatics 30:1031–1033. https://doi.org/10.1093/bioinformatics/btt736.
5. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M. 2009. The NIH Human Microbiome Project. Genome Res 19:2317–2323. https://doi.org/10.1101/gr.096651.109.
6. Dominguez-Bello MG, De Jesus-Laboy KM, Shen N, Cox LM, Amir A, Gonzalez A, Bokulich NA, Song SJ, Hoashi M, Rivera-Vinas JI, Mendez K, Knight R, Clemente JC. 2016. Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. Nat Med 22:250–253. https://doi.org/10.1038/nm.4039.
7. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621.
8. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, DeRight Goldasich L, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, Thompson LR, Tripathi A, Vázquez-Baeza Y, Vrbanac A, Wischmeyer P, Wolfe E, Zhu Q, The American Gut Consortium, Knight R. 2018. American Gut: an open platform for citizen science microbiome research. mSystems 3:e00031-18. https://doi.org/10.1128/mSystems.00031-18.
9. Kyrpides NC, Eloe-Fadrosh EA, Ivanova NN. 2016. Microbiome data science: understanding our microbial planet. Trends Microbiol 24:425–427. https://doi.org/10.1016/j.tim.2016.02.011.
10. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, Gibbons SM, Larsen P, Shogan BD, Weiss S, Metcalf JL, Ursell LK, Vázquez-Baeza Y, Van Treuren W, Hasan NA, Gibson MK, Colwell R, Dantas G, Knight R, Gilbert JA. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science 345:1048–1052. https://doi.org/10.1126/science.1254529.
11. Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, Fontana L, Henrissat B, Knight R, Gordon JI. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332:970–974. https://doi.org/10.1126/science.1198719.
12. Thomas T, Moitinho-Silva L, Lurgi M, Björk JR, Easson C, Astudillo-García C, Olson JB, Erwin PM, López-Legentil S, Luter H, Chaves-Fonnegra A, Costa R, Schupp PJ, Steindler L, Erpenbeck D, Gilbert J, Knight R, Ackermann G, Victor Lopez J, Taylor MW, Thacker RW, Montoya JM, Hentschel U, Webster NS. 2016. Diversity, structure and convergent evolution of the global sponge microbiome. Nat Commun 7:11870. https://doi.org/10.1038/ncomms11870.
13. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–821. https://doi.org/10.1038/nbt.2676.
14. Jing G, Sun Z, Wang H, Gong Y, Huang S, Ning K, Xu J, Su X. 2017. Parallel-META 3: comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. Sci Rep 7:40371. https://doi.org/10.1038/srep40371.
15. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618. https://doi.org/10.1038/ismej.2011.139.
16. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122. https://doi.org/10.1093/nar/gkr1044.